Chapter 2

EXAMINATION OF KNOWN REAL SITES (FROM THE TRANSFAC DATABASE)

2.1 INTRODUCTION

An initial objective was to build up a collection of transcription factor binding sites (TFBSs) that were conserved, and a collection of TFBSs that were clearly non-conserved (diverged), so that possible differences between the two could be investigated.

A "completely bioinformatics" method would use bioinformatics both to discover TFBSs, and to determine whether they were conserved. This method was not used because bioinformatic TFBS searchers produce considerable numbers of false matches. Such an approach has, however, been tried by other authors - (Dermitzakis et al., 2003) did so for fruit flies; they found that many predicted TFBSs were not conserved between species, whereas nearly all the experimentally verified TFBSs were conserved between species. It is somewhat worrying that predicted TFBSs gave a different results to experimentally verified TFBSs. So the present project did not rely on computer-predicted TFBSs.

Instead, the analysis in this thesis started with a list of TFBSs for which there was prior experimental evidence; the TransFac database (Matys et al., 2006) was used. Early versions of the method would retrieve data, align sequences from different species, work out the binding score for possible orthologous TFBSs, then print out their sequences, binding scores, information about the alignment, and various other details, so that a decision could be made manually on whether a TFBS was conserved or not. This was used for a few transcription factors, and the system was improved in the light of what was found during the manual decision-making. Eventually, the completely automated system described here was produced.

2.2 METHODS

2.2.1 Determining whether a TFBS is conserved or not: Outline of method

The computer procedure for determining whether a TFBS is conserved was of considerable complexity; therefore, a general outline will be given here, in addition to the more detailed description later.

For each transcription factor (TF), the process started by retrieving a list of TFBSs from the TransFac database. For a particular TFBS, the database would provide the name of the gene it controlled, the species it was discovered in, and the DNA sequence of the TFBS. Using this gene name, orthologous genes in other vertebrate species were found, using either the HomoloGene or Santa Cruz websites (Wheeler et al., 2006), (Kent et al., 2002), (Kent, 2002). For each species, at least 5,000 bases of upstream sequence was requested from the Santa Cruz genome website (http://genome.ucsc.edu/). The species used were human, rhesus, dog, cow, mouse, rat, opossum and chicken.

For one of these species, we would know that the TFBS exists (from the TransFac database); in that species, the exact location of the TFBS in the

genome was found by searching upstream of the gene for a *perfect match* to the DNA sequence of that particular TFBS (as given by TransFac).

Using BlastZ (Schwartz et al., 2003), this upstream sequence was aligned pairwise with the corresponding sequence from the species that contained the known TFBS. If this alignment was successful, then the DNA sequence of a "possible homologous TFBS" was obtained - that is, the sequence corresponding exactly to the known TFBS. However, if this alignment was not obtained, a conclusion of "uncertain" would be given. This was because failure to obtain alignment, though it might be due to a regulatory region that was not conserved, might also be due to a technical problem, such as

- a gene had been wrongly identified as an orthologue;
- the available mRNA sequence was missing exon 1, so the transcription start site was wrongly identified, and hence wrong upstream sequence was used;
- the available genome sequence was missing sequence in the upstream region;
- the upstream sequence was conserved, but not strongly enough to be detected by the alignment program (there seems to be a considerable risk that an alignment program will fail to detect orthology in a human-mouse comparison (Margulies et al., 2007), see page 57);

Thus, the method ensured that these problems would cause a conclusion of "uncertain", and not an incorrect conclusion of "non-conserved".

Provided the "possible homologous TFBS" was obtained, it was then necessary to decide if it would bind the transcription factor. In principle, this could be done either by comparing the sequence with a consensus sequence, or else by scoring it against a "position weight matrix" (PWM). This research used PWMs from the TRANSFAC database - a different PWM for each transcription factor - those used being based on known TFBSs that exist in a genome. Comparing a possible homologous TFBS to a PWM produced

a "score" between 0 and 1, indicating how closely the sequence matched the sequences known to bind the TF. Hence, a low score would indicate that the "possible homologous sequence" could not, in fact, bind the TF. The TFBS would be classified as "diverged" if it met both the following criteria: (i) the score was below a threshold; (ii) the score was substantially less than the score of the known TFBS. If it met one but not both of these criteria, a conclusion of "uncertain" would be given. If the site met neither criterion, it would be labelled a "conserved" TFBS. The numerical values used in these criteria were set using the scores of the known TFBSs upon which the PWM was based: for criterion (i), the threshold was set as the 20th percentile score minus one standard deviation; for criterion (ii), "substantially less" was set as twice the standard deviation.

Thus, for each known TFBS, the process just described would cause a label of "conserved" or "diverged" or "uncertain" to be given to each of the seven vertebrate species that were being used as comparisons. For example, a TFBS that is known to exist in humans (from experimental evidence) might be "conserved" in chimp, but the same TFBS might be "diverged" in mice.

2.2.2 Determining whether a TFBS is conserved or not: Detailed method: the PWMs

Choosing the PWMs

In choosing which matrices to use, one distinction was between "selected" matrices, which were based on sequences found by *in-vitro* selection, and "compiled" matrices, based on genomic TFBSs. The former method uses random oligonucleotides which contain a vast number of short DNA sequences; these migrate *in-vitro*, but a particular TF is present and will bind some of the oligonucleotides, thus slowing their migration. By this means, the oligos that have a high binding affinity are separated from the rest. In contrast, the "compiled matrices" are based on TFBSs that are believed to exist in the genome, generally based on experimentally verified TFBS sequences

(although the type of its evidence and its reliability may vary).

The "compiled" matrices were assumed to be more realistic, largely because of concern that the selected-oligo procedure might sometimes be too strict. As an example, consider the "selected" TransFac matrix M00034 and the "compiled" matrix M00272, both of which are intended to represent TFBSs bound by p53. Looking at the entries, it is evident that the sequences on which M00272 was based differed considerably from the perfect consensus sequence; the sequences on which M00034 was based showed much less variation. Perhaps this is because the selected-oligo procedure used so many cycles of purification that the only sequences obtained were those with the highest possible binding affinity. In contrast, a study of genomic TFBSs found that some showed weaker *in-vitro* binding than others (Tronche et al., 1997). So it was decided to use only matrices compiled from genomic TFBSs.

By deciding to use only matrices compiled from genomic sequences, the number of possible sources of data were reduced. The JASPAR database contains a large number of PWMs, so for many projects, it is possible to use JASPAR as a source of PWMs, instead of TransFac. However, at the start of this project, most PWMs in JASPAR were based on *in-vitro* selection. Only 20 PWMs in JASPAR were complied from genomic sequences (Sandelin et al., 2004), whereas far more PWMs of this type were available from TransFac. Hence JASPAR was not used.

During the initial development of the system, matrices were chosen in a somewhat *ad hoc* manner (these TFs are shown in the "initially selected" list of table 2.1). Subsequently, however, they were chosen from a list of all compiled TransFac matrices, listed in order of the number of TFBSs they were based on. No consideration was given to TFs which only had matrices based on less than 15 TFBSs.

In many cases, a particular TF is represented by more than one matrix in TransFac. Consequently, if every matrix in TransFac was used, then the list of matrices would contain many TFs in duplicate (or triplicate, etc), which was considered undesirable. Therefore, a great many matrices had to be rejected simply to avoid duplicating TFs. In some cases, two matrices representing the

same TF would be quite similar, so the exact choice of matrix was arbitrary. To avoid arbitrary choices as much as possible, a list of points taken into consideration when choosing PWMSs is shown in tables 2.1-2.2.

Some matrices are compiled from individual TFBSs by the TransFac team; however, some TransFac matrices are based on matrices or tables that were first published in the scientific literature. Not surprisingly, there are some matrices from the literature that had been compiled on an unusual basis - for instance, M00116 is based entirely on TFBSs from genes "expressed in the liver but not increasing during acute phase response", and M00117 is based on genes "expressed during acute phase response". No other TransFac matrices contain the words "acute phase response". It was impossible to anticipate all such novelties before looking at the matrices. Nor was it possible to produce (before looking at the matrices) a formula defining which novelties were serious enough to require that the matrix be rejected. Instead, each TF on the list (of "compiled" matrices with at least 15 TFBSs) was considered manually. This was to decide which of the matrices would be used to represent that TF; many of the criteria in tables 2.1-2.2 were developed in the light of this experience.

The "initally selected" TFs were re-examined at the same time that the other TFs were being selected, to check that they met the same criteria that the other matrices had to meet. For two of them (HNF-1 and NF-1) the matrix used was changed, as in each case the matrix originally used would not have been chosen during the later procedure.

The manual examination was also to decide precisely what symbol would be used. For example, the symbols POU2F1 and Oct-1 are synonyms for the same TF; querying TransFac with "POU2F1" produced far more TFBSs than querying with "Oct-1"; therefore POU2F1 was used as the name for both the PWM and the TFBSs (even though TransFac named that PWM after Oct).

If all available PWMs were used in the exploratory stages of the study, it would be difficult to verify any results obtained. Therefore, a number of PWMs were put on a "reserve list". These were not used during the main analysis, and used only during verification of a small number of key results.

To select these, once two PWMs had been chosen as suitable for use in analysis, they were treated as a "pair", and by tossing a coin one PWM was selected to be added to the main analysis, and the other selected to be added to the reserve list. HNF-4 was originally included in the reserve list but, before the reserve list was analysed, HNF-4 was removed as chip-chip data

Table 2.1: Points considered when choosing PWMs

- Most TranFac matrix entries include a list of the sequences from which they were compiled, but some do not; the latter were not used since the sequences were needed to estimate score averages and standard deviations, etc.
- PWMs that are only intended to represent half-sites were considered acceptable on the grounds that, if a TFBS became useless, this would be evident from the accumulation of mutations even if only half the TFBS was observed (though more evolutionary time might be required before it was evident, than if the whole TFBS was being observed).
- Consideration was given to excluding PWMs with a high frequecy of false matches (i.e., matches in random sequence), but it was decided to include these as the analysis contained safeguards against false matches.
- Sometimes TranFac offers a choice of PWMs for a particular TF, one being based on all available TFBSs and the other(s) based only on TFBSs of a higher experimental quality. The higher quality PWM would be chosen, provided it was based on enough TFBSs.
- Some PWMs are based on a special subset of TFBSs - for instance, TFBSs occupied during acute phase in the liver - but, in such cases, it was preferred to use a general PWM for the same TF.
- It was occasionally necessary to change the TF name for a PWM when it became clear that most of the TFBS entries used a different name for the same TF (for example, Oct-1/POU2F1, as explained in the main text).
- Sometimes a PWM represented a family of similar TFs (eg the HNF-3 family) but another PWM represented a single TF in the same family; those were chosen on a case-by-case basis.

Table 2.2: Points considered when choosing PWMs, continued

- The length of PWMs (i.e. number of bases) could differ even if they represented the same TF; then very short PWMs (6 bases long) were avoided, and moderately short PWMs (9 bases long) tended to be avoided though this was regarded as unimportant compared with other items on this list.
- In one case PWMs were available that represented the same TFs but with different lengths of spacer between the half-sites; a PWM giving only a half-site was preferred, as then there was no risk of using a PWM which had the wrong spacer length for a particular TFBS. Whilst this would be insensitive to changes that only affected one half of a TFBS, it was argued that once a TFBS went out of use, both halves would tend to accumulate mutations, and so analysis of the half-site alone would often detect this.
- More than one PWM represented a TF (or family of TFs) that often heterodimerise with RXR, and since RXR was already represented by another PWM, including both PWMs could cause the heterodimer sites to be analysed twice; one such PWM was rejected as nearly all the relevant TFBSs were heterodimers with RXR, but another was accepted as half the relevant TFBSs were not shown as involving RXR.

(Odom et al., 2004) became available for HNF-4, so a separate analysis was made for HNF-4 (see page 136). Tables 2.3 and 2.4 show the two lists.

Scoring matches against the PWMs

A special Perl module determined how well a DNA sequence matched a particular PWM. The original version of this module was written during the author's MSc project at the University of Exeter, was subsequently improved whilst the author was a research assistant at the University of Exeter, and was not altered much during the course of the present PhD project.

A frequency matrix could be accepted as input, and would be converted to a "matrix of proportions". This convertion used a method that is mathemat-

Table 2.3: TFs and matrices used - main list

MAIN LIST	
Initially selected	
HNF-1	M00790
CREB	M00916
NF-1	M00806
NF-AT	M00935
NF-Y	M00775
NF-kappaB	M00054
SRF	M00810
TTF-1	M00794
ETS	M00771
Randomly selected	
HNF-3alpha	M00724
HNF-3beta	M00791
HNF-3gamma	M00791
SP1	M00933
C/EBP	M00912
p53	M00761
GR	M00921
USF	M00796
PXR [|CAR|LXR|FXR]	M00964 Not used - nearly all the relevant TFBSs were for RXR heterodimers, and RXR is listed (see next table)
STAT	M00777
Crx	M00623
c-Myb	M00773
AP-2	M00800
YY1	M00793
Pax	M00808
PR\W	M00960
POU1F1	M00744
The symbols shown are those used to query TransFac.	

ically equivalent to Bucher's equation (see page 30) with s=1%, $c_i = 0$, and the "background" frequency assumed to be 25% for each base. However, the actual implementation omitted the logarithms, but used the mathematically equivalent system of multiplying elements, rather than adding the logarithms of those elements. Thus a "matrix of proportions" was produced in which each element equalled

$$\left(n_{bi} / \sum_{b=a,c,g,t} n_{bi}\right) + s/100 \tag{2.1}$$

.

To calculate a binding score for a DNA sequence (where the sequence was the same length as the PWM), the following procedure was used. Table 2.5 shows a worked example of this, using a PWM which, for simplicity, is rather short (5 rows) compared with real PWMs. The general procedure was that each base in the sequence was given a score equal to the relevant cell in the matrix of proportions; these were multiplied by each other to give an overall multiplicative score for the sequence. The logarithm of the multiplicative

Table 2.4: TFs and matrices used - reserve list

RESERVE LIST	
Randomly selected	
E2F	M00919
AP-1	M00924
GATA	M00789
POU2F1	M00930
ER	M00959
T3R [RAR|RXR]	M00963
HNF-4 [HNF-4alpha|HNF-4alpha1]	M00638 Not used
AR	M00962
AhR	M00778
IRF [ICSBP]	M00772
HIF-1	M00797
c-Myc	M00799
VDR	M00961
MyoD	M00929
AML	M00769
The symbols shown are those used to query TransFac.	

score was taken, and put through a linear transformation so that the highest possible score was 1 and the lowest possible score was 0. The result of this was called the binding score.

The problem of cells with a count of zero was addressed, not by pseudocounts, but by using the method (Tsunoda and Takagi, 1999) of having a smoothing parameter of 0.01.

A long DNA sequence could be provided as input and the module would return an array of objects, where the first object in the array contained details of the best (ie highest-scoring) match to the PWM, the second object contained details of the second-best match, etc. The input sequence could contain "n" and "-"; "n"s would be scored using the average of the four cells in the matrix, "-" would be scored using a "gapScore" which needed to be small and so was arbitrarily set equal to the smoothing parameter (unless the latter was zero and the input matrix included text stating that the matrix already included the smoothing parameter, in which case it was set equal to 0.01). Apart from these, and the four DNA bases, any other character in the sequence would stop the program.

Reverse complement matches were treated as equally valid as forward matches.

Using shuffled matrices to obtain p-values

False match rates were estimated using "shuffled matrices"; each of these was produced by taking a real PWM and shuffling the rows at random. (The matrix format had four columns corresponding to A, C, G and T, so each row corresponds to a different position of the DNA sequence). Each would be used to search real DNA sequence for matches that had a binding-score larger than a particular value (where the particular value was a binding-score whose statistical importance we needed to assess). The exact way these were used will be described in subsequent flowcharts, etc.

Each shuffled matrix would have several properties identical to the real PWM from which it was created, such as the GC content; the "information content" (Schneider et al., 1986) would also be the same. It was assumed, therefore,

Table 2.5: Worked example of using a PWM to score a DNA sequence

An example of a frequency matrix is shown below. This will have been compiled from the sequences of 71 TFBSs that bind the transcription factor we are studying, each TFBS being 5bp long. For example, nearly all these sequences started with the base "A"; more precisely, 64 out of 71 sequences started "A", so the 64 is entered in the first row of the matrix:

A	C	G	T
64	1	2	4
57	3	10	1
60	2	8	1
4	4	62	1
6	1	35	29

This is converted to a matrix of proportions, by dividing each frequency by the row total, then adding a smoothing parameter of 0.01. For example, the top left cell becomes 64/(64+1+2+4) + 0.01 = 0.911. This gives:

A	C	G	T
0.911	0.024	0.038	0.066
0.813	0.052	0.151	0.024
0.855	0.038	0.123	0.024
0.066	0.066	0.883	0.024
0.095	0.024	0.503	0.418

Let us use this to score a sequence, for example CAAGA. The first letter is C, so we take the number in the C column of the 1st row; then we take the number in the A column of the 2nd row, etc, divide each number by 0.25, and multiply them all together:

$$0.038/0.25 * 0.813/0.25 * 0.855/0.25 * 0.883/0.25 * 0.095/0.25 = 2.27$$

This particular sequence has given 2.27; but the highest possible value would be 288 (produced by AAAGG), and the lowest possible would be 0.00000815 (produced by CTTTC). The final score for CAAGA is obtained by taking the natural logarithms of these and comparing as follows:

$$BindingScore = \frac{log(2.27) - log(0.00000815)}{(log(288) - log(0.00000815)} = 0.805$$

that the false-match rate of a shuffled matrix was similar to that of the real PWM from which it was created - however it is not claimed that they would be perfectly identical.

The shuffled-matrix method has been used for some time (Tronche et al., 1997) and was implemented in a Perl module during the author's time at the University of Exeter (Lockwood and Frayling, 2003), this module being re-used in the current project. Other researchers have also used shuffled PWMs to search genomic sequence, for instance Lander's group (Xie et al., 2007). The JASPAR web database has recently included a tool for generating shuffled matrices (Bryne et al., 2008).

2.2.3 Determining whether a TFBS is conserved or not: Detailed method: Obtaining Data

Setting the threshold score

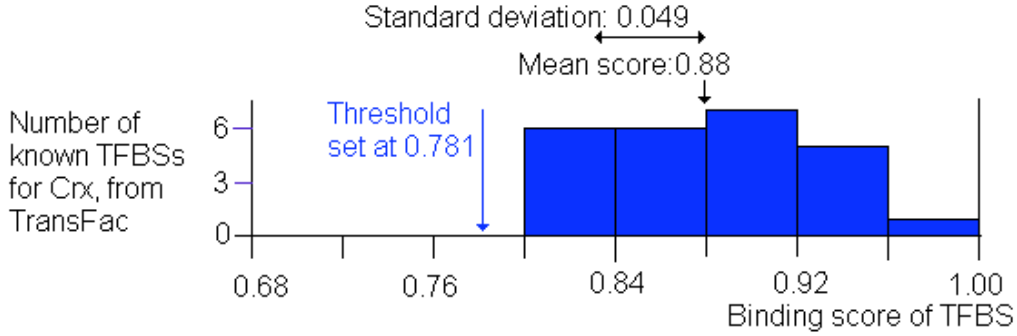
The binding properties of each particular transcription factor were described by a position weight matrix (PWM), which could be used to "score" any DNA sequence. The importance of a particular score was assessed by comparing it with the scores obtained by actual TFBSs (these were the same TFBSs as the PWM was compiled from). The PWM was used to give a score to each sequence, and from these scores were estimated the mean score / standard deviation / 10th percentile, 20th percentile, median, 80th percentile, and 90th percentile scores. The latter were estimated straight from the data (for example, if there were 30 sites, the 20th percentile score was taken to be the average of the 6th and 7th lowest scores) - no attempt was made to fit a frequency distribution curve. As an example, Fig 2.1 shows a histogram of binding scores for TFBSs that bind Crx.

These data were used to set the "threshold score"; the aim of this was that any sequence with a score lower than this threshold would be regarded as unlikely to be a true TFBS.

In principle, one might set the threshold score as being the lowest of the scores obtained by real TFBSs in the database, as this would ensure that a sequence that obtained a lower score than any real TFBS would not be regarded as a TFBS. But that would have the disadvantage that the threshold score used could be drastically affected by a single incorrect TFBS in the database, or by a single case where imperfections in the PWM caused an unusually low score. These disadvantages could be overcome by setting the threshold to be the 5th percentile score. However, the 5th percentile score will not be known with accuracy when, typically, 30 TFBSs or fewer were available for a particular transcription factor. Therefore the threshold was set at the 20th percentile score minus one standard deviation (on the grounds that, if the scores were Normally distributed, this would be similar to the 5 percentile score).

On Fig 2.1 the threshold score is marked for the TF Crx. In this case the threshold score is just below the score of all the known TFBSs, but in general it is possible to have a few known TFBSs with scores below the threshold.

Figure 2.1: Histogram of binding scores for Crx TFBSs
This uses the TFBSs on which the PWM for Crx was based.



Why use threshold scores that differ from those used in other studies?

The threshold scores just described are different from those used by other studies that employ PWMs. The reason for this is as follows. The main use of PWMs is to find possible TFBSs in DNA sequences, so the threshold scores are usually set with this in mind. Because searches of this kind produce large numbers of false matches, researchers try to prevent this by setting the threshold score as high as possible. For instance, the threshold score may be set so that 50% of real TFBSs are not recognised (Goessling et al., 2001). Obviously, this will cause 50% of genuine TFBSs to be overlooked during the search for TFBSs, but this is regarded as a price worth paying to avoid vast numbers of false matches - a reasonable attitude in the context of TFBS searching.

However, the present project has the quite different purpose of identifying diverged TFBSs, and adopting the same threshold scores would be incautious.

As an example, suppose a Crx TFBS in human had a binding score of 0.94 and there was an orthologous Crx TFBS in mouse with a binding score of 0.84. Both these scores are in the normal range of Crx scores (see figure 2.1), so this could well be a conserved TFBS. However, if we use the same threshold as projects that search for TFBSs, and use the 50% (median) score as a threshold, then since the human TFBS is above this score and the mouse TFBS is below it, it *appears* that the TFBS exists in humans but not mice, and thus *appears* to be an example of a TFBS that has diverged. But that would be an incautious conclusion, since (as just noted) this could in fact be a conserved TFBS.

As this example shows, for the purposes of the current project, it would be incautious to assume that a sequence was not a TFBS because its binding score was below the median binding score. Yet the same assumption could be quite appropriate for a project that was searching for new TFBSs. The current project has to avoid the problem of conserved TFBSs being mistakenly classified as diverged TFBSs, whereas a TFBS search has to avoid the

problem of generating vast number of false matches.

Because of this difference, it would not be appropriate to use threshold scores used elsewhere, as these have generally been chosen for the purpose of searching sequences for new TFBSs.

Obtaining a list of TFBSs for a particular transcription factor

For a given species and TF, a list of TFBSs was obtained from the SITE table of the TransFac Professional database. Generally, versions 10.2/10.3 were used, but some results from earlier verions will be noted later. The "TFBS" Perl package (Lenhard and Wasserman, 2002) was used to store this information within the program.

Identity of a gene

The acronym of each gene was checked by querying if it was in the "symbol" field of the database at the HUGO human gene nomenclature website (www.gene.ucl.ac.uk/cgi-bin/nomenclature/searchgenes.pl)
(Wain et al., 2004), and if that was unsuccessful, the "symb_prev_and_alias" field. Non-standard gene names would be converted to standard gene names. If these searches produced no record, or more than one record, analysis of that TFBS was abandoned.

Obtaining the homologous genes in other species

Human/mouse/rat homologues were obtained by contacting NCBI's HomoloGene website (www.ncbi.nlm.nih.gov/HomoloGene) (Wheeler et al., 2006). This database contains homologues generated by computer comparison of sequences, and is not curated.

The software queried HomoloGene using the gene symbol (or, if available, the RefSeq accession number (Pruitt et al., 2005)). The reply would be a list of HomoloGene entries, only one of which was wanted. The entry with

a name identical to the gene symbol would be used (if none had an identical name, the first entry on the list would be used).

A second request to HomoloGene obtained the actual entry. Originally, data were taken only for human/mouse/rat homologous genes that had an identical symbol to the gene being investigated. However, it was realised that this could cause genes to be rejected unnecessarily due to small differences in gene symbols. For example, humans have one insulin gene (INS) whereas mice have two (Ins1 & Ins2). Therefore, the requirement was relaxed so that the two symbols had to be identical except that a single extra character on the end of one symbol was permitted. (A related exercise would be to find if a TFBS had been conserved in one paralogue but not the other, but that was regarded as being outside the scope of this study.) The data obtained would be the identity number of a protein, so a further request to the NCBI website was used to obtain the RefSeq accession number for that gene.

However, as HomoloGene did not cover all the species that eventually were used in this project, an alternative source of homology information was required. The XenoRefGene system of the Santa Cruz Genome Browser was used (http://genome.cse.ucsc.edu/cgi-bin/hgGateway). This feature of the website does not have its own well-publicised name; here it will be referred to as the "XenoRefGene" system, as that name is used in the CGI requests. The system for querying this by computer will be described shortly, but it will be easier to understand if one first considers how the same task could be done by a casual user visiting the Genome Browser website. Imagine that the casual user first visits the Browser Gateway for a species such as cow (which is not one of the popular model animals) and they query it using a gene symbol, say "GFAP". The website responds by producing a list of non-cow genes which have been mapped to the cow genome. Clicking on one of these produces a display of that part of the cow genome, and the display also shows non-cow genes; clicking on one of these produces a "XenoRefGene" page, which includes a very short table, showing the co-ordinates and direction of the part of the cow genome that is aligned to the non-cow gene.

The computer-driven version of this was essentially the same, although some

additional details had to be explicitly considered. For details of this procedure, see figure 2.2.

If a particular species occurred more than once in the list of homologues, that TFBS would not be analysed as the programs were not designed to deal with such a situation.

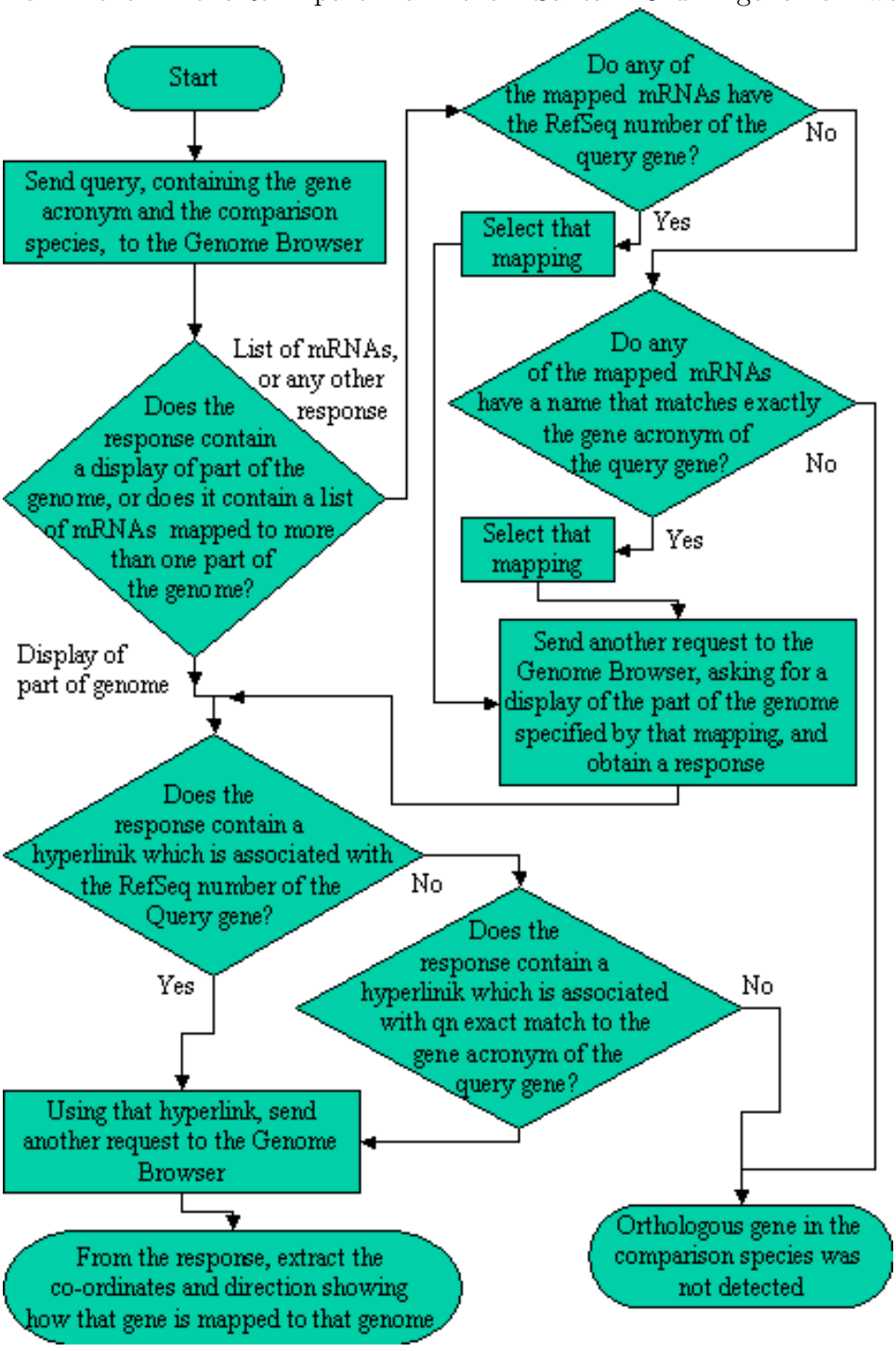
Obtaining DNA sequence near the TFBS

This re-used software I originally wrote for my MSc at the University of Exeter. By this stage in the analysis, we should already have a RefSeq accession number (Pruitt et al., 2005), which refers to the gene controlled by the TFBS. This is the basis for obtaining the required sequences. First, the mRNA sequence of that gene was obtained by contacting the NCBI website (www.ncbi.nlm.nih.gov). Next, genome sequence was obtained from the database (http://genome.cse.ucsc.edu) at Santz Cruz. For this, the first 50 bases of the mRNA sequence was used as a query to their "BLAT searcher" to find the exact co-ordinate of the start of the mRNA in the genome. (This step was not required if a homologue had been found using the Santa Cruz website, since, then, the exact coordinate was already available.) DNA sequence was downloaded based on the coordinate. The sequence retrieved was usually 5k upstream to 10k downstream of the transcription start site, although extra upstream sequence could be retrieved if the TFBS was expected (from the coordinate given by TransFac) to be far upstream. The downstream sequence was retrieved so that protein-coding sequence was available, for determining the synonymous mutation rate.

Because of difficulties in keeping the above system working, towards the end of the project human, mouse and rat genome sequence was instead obtained from Ensembl (Birney et al., 2006). Each gene was located using a file that had been downloaded from Ensembl40; this file contained gene names, RefSeq numbers, and co-ordinates that showed where each gene was located on the genome. Starting with a RefSeq accession number of a gene (or if that was not available, starting with a gene name (acronym)), the co-ordinates of

Figure 2.2: Flowchart showing how orthologous genes were obtained from XenoRef

Flowchart showing how orthologous genes were obtained from the XenoRef part of the Santa Cruz genome website.



that gene were looked up within the file. Using these co-ordinates, a request was sent to the Ensembl website, asking for genomic sequence near the start of the gene. Once this genomic sequence had been obtained, the position of the TSS was estimated by searching the genomic sequence for a match to the first 50 bases of mRNA sequence. Since both sequences came from the same species, one would generally expect to obtain a perfect match. However, surprisingly, there were a number of cases where a perfect match could not be obtained, but an imperfect match could be (eg RefSeq entries NM_010591 and NM_008872); this might be due to polymorphisms or mistakes in sequencing. Therefore a match was required with 10 mismatches or fewer, and no gaps, to the first 50 bases of mRNA, and if that was obtained, then it was assumed to give the location of the TSS of that gene.

Local storage of data

Originally, the programs retrieved information from distant websites as they went along. This, however, meant that those websites would receive the same query a number of times. To avoid this, was written to save the response from each query locally, and retrieve that if the query was made again. Where this might have meant using out-of-date data, it was arranged that locally-stored data would not be used if more than 3 months old.

2.2.4 Determining whether a TFBS is conserved or not: Detailed method: Analysing Data

Locating the precise TFBS

The co-ordinates obtained from TransFac were not relied on (for reasons that will be evident later when figure 2.11 is considered). Instead, the TFBS was located by examining the sequence upstream of the gene specified in the TransFac entry, requiring a *perfect match* to the sequence of the TFBS given by TransFac. The search was from c_{search} relative to the transcription start site, to the transcription start site, where c_{search} was given by:

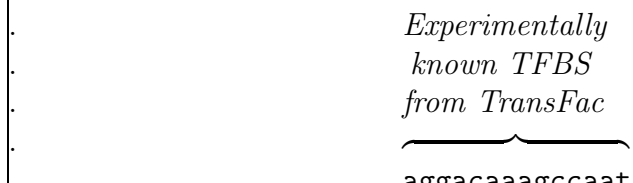
Table 2.6: Example of locating the precise TFBS

An actual example showing how the precise TFBS would be located.

In TransFac, entry R08083 of the "Site" table states a TFBS binding HNF-1 is located -67 relative to the human Fib-A (FGA) gene, and gives its sequence as aggacaaagccaat. Human genome sequence upstream of this gene was retrieved as described earlier, and used as follows.

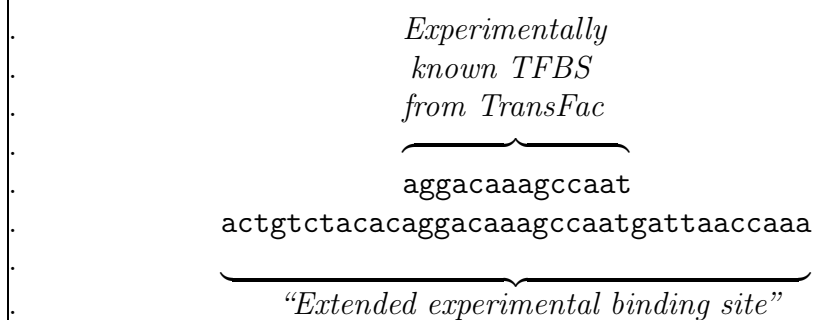
The location of the upstream end of the TFBS was given as -67, so the search region started 2*(-67)-400 = -534 bases upstream of the gene.

The genomic sequence from -534 to 0 (relative to the start of the FGA gene) was searched for an *exact match* to the aggacaaagccaat sequence given by TransFac. If no exact match was found, analysis of this TFBS would have been abandoned. In this case, the exact match was found:



Human genome sequence upstream of FGA

The sequence from TransFac, aggacaaagccaat, was shorter than the PWM used to describe HNF-1 binding sites (14 bases, whereas the PWM described an 18 base sequence). Thus the difference is 4 bases. Therefore the sequence was extended by 4 bases on either side. In addition, every TFBS was extended by 7 bases on either side, to produce the "extended experimental binding sequence"; thus in this example, a total of 11 bases (=7+4) on either side was added:

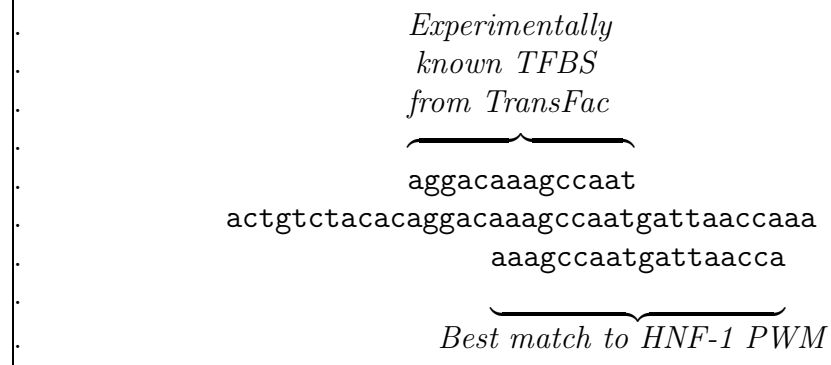


(continued...)

Table 2.7: Example of locating the precise TFBS, cont.

(...example continued from table 2.6)

The "extended experimental binding site" was then searched to find the best match to the HNF-1 PWM. In later analysis, this would be regarded as the "true" TFBS. The match-score was 0.896.



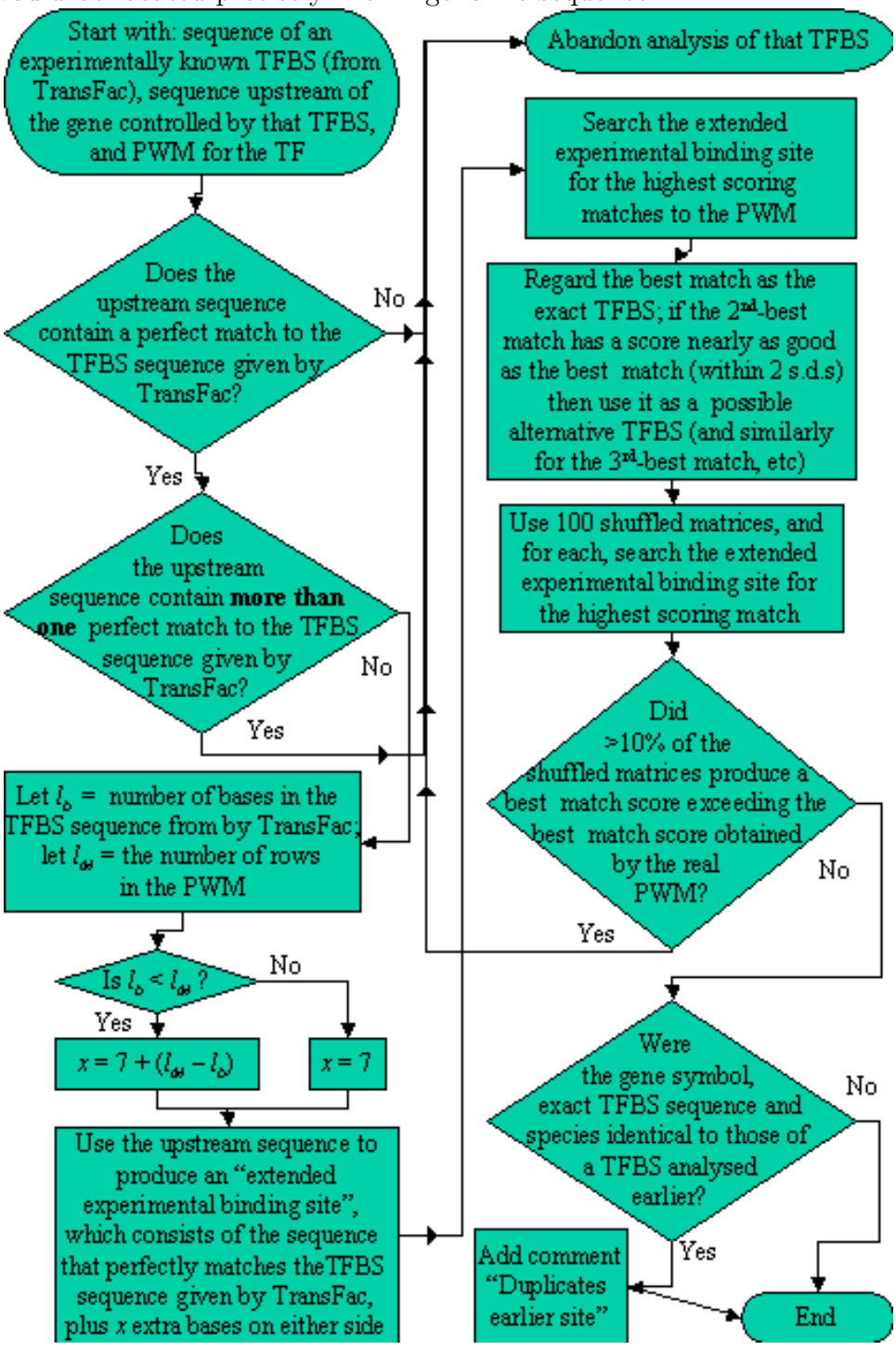
A shuffled version of the PWM was then used to search the extended experimental binding site for the best match. This was repeated, until 100 shuffled PWMs had been used. None of these produced a match with a score as high as 0.896. This suggests that a match score of 0.896 was very unlikely to have been produced by chance. (Conversely, if more than 10 shuffled PWMs had produced a score greater than 0.896, analysis of this TFBS would have been abandoned).

The upstream end of the best match had a coordinate of -57, if measured using the genomic sequence, relative to the start of the FGA gene (as represented by RefSeq entry NM_000508).

COMMENTS: It might be wondered whether too many bases were used when extending the TranFac sequence. However, if we used the bare minimum extra bases to make it the same length as the PWM (2 on each side), the resulting sequence (acaggacaaagccaatga) obtains a match-score of 0.487, which is too low to suggest a plausible HNF-1 site (90% of known HNF-1 TFBSs scored more than 0.66). This suggests the sequence given by TransFac was both too short, and also off-centre relative to the true TFBS.

It will also be noticed that the location given by TransFac, -67, differs from the location estimated from the genomic sequence, -57. The diagram above would lead one to expect a discrepancy of 5 bases. This partly, but not wholely explains the actual discrepancy.

Figure 2.3: Flowchart showing how a TFBS would be located precisely Flowchart showing how a TFBS, which was already known from TransFac, would be located precisely within genomic sequence.



$c_{search} = 2 * c_{TFBS} - 400$

where c_{TFBS} was the co-ordinate of the upstream end of the TFBS, as given by TransFac (unless this co-ordinate was greater than zero, when a slightly different formula was used.) The 400 was chosen after early experience produced a few cases where a span of 400bp was necessary to find the TFBS.

Very often, the individual TFBS sequence from TransFac is shorter or longer than the PWM, or appears to overlap the PWM rather than coinciding with it completely. In these cases the exact sequence and location of the TFBS, which formed the basis for subsequent analysis, was not the same as that given by TranFac. Tables 2.6-2.7 give an example where this was done, whilst figure 2.3 describes the procedure in more general terms.

Note that in some cases, this procedure could lead to more than one location being a possible place for the TFBS - that is, there was uncertainty as to the exact location of the TFBS. Where it was uncertain which of two possible locations was the true TFBS, details of both would be stored and later analysed; the one with the lower match-score being called the "2nd-best match" in figure 2.3. The underlying principle was that, if it was not certain which location was correct, then if analysis of each lead to opposite conclusions, neither conclusion should be relied on. (In practise, there have been quite a number of cases where both the best and 2nd-best sites are conserved; thus we could conclude that the true TFBS is conserved even though we were uncertain about the exact location of the TFBS).

To ensure a very low probability of obtaining a match by chance, only TransFac sequences at least 10 bases long were analysed (a considerable number of TFBSs in TransFac are recorded as only a 6-base sequence). If no match was obtained, the analysis of the binding site was abandoned with the message "Experimental site seq not found in genomic sequence". If two matches were obtained, analysis was also abandoned. Note that this adds a degree of robustness to the procedure - if the wrong official gene symbol is used, or the co-ordinates are grossly in error, then the analysis will very likely be abandoned at this stage (because the experimental TFBS sequence could not be found in the genomic sequence), and not continue with the risk of producing

misleading results.

To check that the TFBS was a good match to the PWM, a "shuffled matrix" search of the extended experimental binding site was also carried out, as detailed in figure 2.3. This check could cause analysis of that TFBS to be abandoned. This might be because the experimentally determined binding sequence was incorrect, or because the transcription factor bound that TFBS in a way that was not well represented by the PWM - both of these being a good reason to abandon the analysis.

Alignment of sequences

The homologous sequences from the different species were aligned using BlastZ (Schwartz et al., 2003), an alignment program specially designed to align genomic sequences of human and mouse (see page 58 for additional details about BlastZ). When developing the method, initially the ClustalW (Thompson et al., 1994) system was tried, but rejected because it would align any sequences it was given - a feature that would have a very undesirable effect if it was given two unrelated sequences by mistake, since the known TFBS in one species would be aligned against an unrelated sequence, and give the appearance of not being conserved. Therefore, it was thought better to use an alignment program capable of giving the result "No significant similarity found". Blast2 (Tatusova and Madden, 1999) was used for a while and, on a trial basis, T-Coffee (Notredame et al., 2000) was used to re-align the output from Blast2; T-Coffee would often insert a gap in the TFBS when Blast2 had not. Subsequently, Simon Hubbard (personal communication) suggested BlastZ would be more sensitive at detecting alignments in the less well-conserved parts of the genome.

Before aligning, sequence downstream of the TSS was removed. This was to ensure that none of the alignment included any protein-coding sequence, as it was intended to use statistics about the alignment as a description of the regulatory region. All the upstream sequence retrieved (usually 5k, but more if c_{search} was that large) was submitted to the alignment program.

The alignments were pairwise, where one of the sequences was upstream sequence that included the known TFBS, and the other was orthologous sequence from another species.

BlastZ was called with options B=0 C=2 K=2000. K is the alignment score required for an alignment to be produced; as it was important to reject unrelated sequences, a separate test was carried out before choosing to put K=2000 in the main program. The test used the sequences upstream of human genes, which had already been retrieved during the project. A pair of these sequences was chosen at random and submitted to BlastZ. Since the two sequences would nearly always be unrelated, no alignment should have been found. In fact, an alignment was produced for only 30 pairs out of 2000 pairs tried. This suggests that the K=2000 setting would nearly always succeed at preventing an attempt to align unrelated sequences.

Examination of alignments

In the alignments produced, often the sequence containing the known TFBS would be aligned against sequence from the other species. However, it was possible instead for no alignment to be produced at all, in which case an "Uncertain" conclusion would be reached, as in these cases it was impossible to be sure whether a TFBS was conserved or was non-conserved.

Also, an "Uncertain" conclusion would usually be produced if some sequence was aligned, but the known TFBS was outside the aligned sequence. However, there were exceptions to this: (i) if the known TFBS was partly outside the aligned region (by not more than 5 bases), it would be treated as if it was aligned; (ii)A special procedure was followed if the known TFBS was between two aligned regions (the details of this will be given on page 103). This was used if the two aligned regions were in the same order in both species, and provided the sequence between the two aligned regions was of roughly similar length in both species (defined as differing by less than a factor of 2).

Deciding if a TFBS was conserved or not

The software would generate a "comment" at various stages of the analysis, and these comments would later be the basis of producing a single conclusion about whether a particular TFBS was conserved in a particular species.

From the pairwise aligned sequences, a "splice" would (where possible) be extracted, which contained only the known TFBS and the sequence aligned against it. The latter was known as the "possible homologue TFBS". It was scored using the PWM to assess how well it would bind the transcription factor.

A "possible homologue TFBS" would be given a comment of "non-conserved" ("Probably diverged" was the terminology used in the program) if it met both the following criteria: (i) the score was below a threshold (to see how the threshold was set, see page 80); (ii) the score was at least two standard deviations below the score of the known binding TFBS.

For an imaginary example of this, see fig 2.4A. Here a known Crx TFBS in human, with a binding score of 0.91, was aligned against mouse sequence that was found to have a binding score of 0.71. This meets both criteria to be considered "non-conserved". The histogram shows scores of actual TFBSs, and is included in the figure to help the reader judge if the mouse score is too low to be a TFBS.

If the possible homologue met neither criterion, it would be given a comment of "Probably conserved" site (or simply "Conserved" if the score was high enough). Fig 2.4B shows an imaginary example of this.

If the possible homologue met one but not both of these criteria, a comment of "Unclear" would be given; examples are shown in fig 2.4C,D. These figures help illustrate why *two* criteria were thought necessary.

The case in fig 2.4C meets the first criterion to be non-conserved, since the mouse score is just below the threshold, and yet it does not look like a convincing example of a non-conserved TFBS, given that the exact setting of the threshold is somewhat arbitrary. Sensibly, a comment of "uncertain"

is given because it does not meet the second criterion.

Fig 2.4D gives an example where criterion (ii) was met but not criterion (i), again leading to a comment of "uncertain". In contrast, another bioinformatic study of TFBS evolution (Sauer et al., 2006) used only one criterion, the change in binding score; thus in that study, a case similar to fig 2.4D could be classified as a "non-conserved" TFBS. The argument against doing so is that, since the mouse sequence (in this example) has a binding score higher than some known TFBSs, then there should be doubts about assuming that that sequence is not a TFBS.

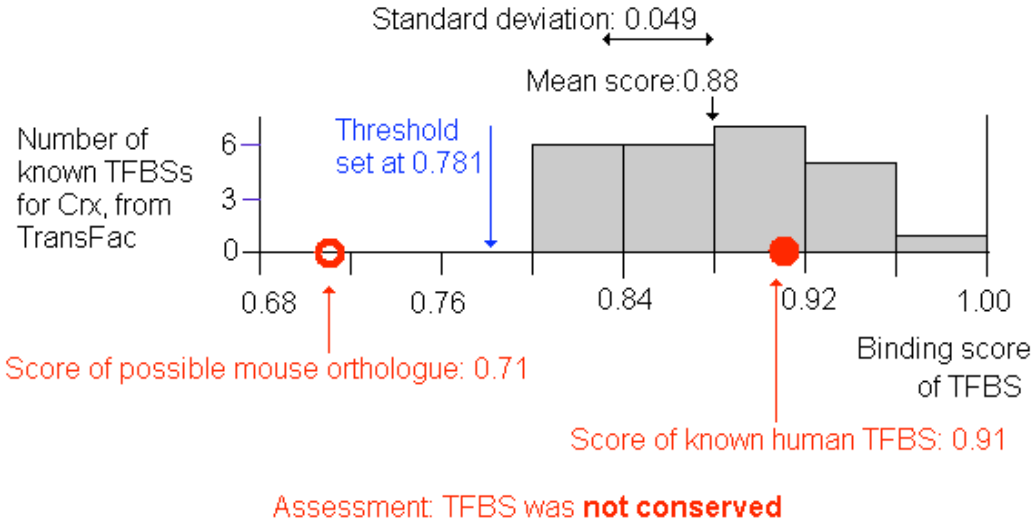
A "Probably diverged strongly" comment would be given if the binding score was so low that it not only met both conditions, but was also closer to the score generated by a random sequence (median value) than to the threshold. However, all the analyses presented treated this the same as "probably diverged".

Deciding if a TFBS was conserved or not: when a gap is in a TFBS

When the possible homologue TFBS contained gaps, the above procedure was applied, but since the PWM-scoring system was set up to give a very low score to a gap, the entire homologue TFBS would receive a low score. This was originally thought justified on the grounds that a deletion event within a TFBS would be likely to destroy its ability to bind. However, it is not certain that every insert/deletion in a TFBS disrupts binding; moreover, it was also thought that the exact position of a gap inserted by an alignment program might be arbitrary, so a gap could be inserted in a TFBS when it could plausibly be inserted instead just outside the TFBS. Table 2.8 shows an actual example where it seems possible that a deletion within a TFBS had little or no effect on binding. Given this, it seems incautious to assume that a gap is proof of a non-conserved TFBS.

To address this problem, the gaps would be removed and the sequence scored using the PWM. Removing the gaps would leave a TFBS that was too short, so extra sequence had to be taken from either side and assumed to be part

Figure 2.4: Criteria for deciding if a TFBS is not conserved
In this example, we consider a TFBS upstream of a particular human gene that is known, from the TransFac database, to bind Crx. Previous stages in the analysis have identified a DNA sequence upstream of the orthologous mouse gene that *might* be an orthologous TFBS. The binding scores of these are used to decide whether this TFBS is conserved or non-conserved, and two criteria (defined in the main text) are used for this purpose. This figure shows four examples. Fig 2.1 is shown as a "grey background", because the criteria used depend on that data.



A (above). The possible mouse orthologue has a binding score which is below the threshold *and* well below the score of the human TFBS (more than two standard deviations below). Thus it meets both criteria for a non-conserved TFBS, and is assessed as *not conserved*.

of the TFBS; and since it was unclear whether the extra sequence should be taken from the left or the right, it was taken from both sides, resulting in a sequence that was longer than required; this was searched using the PWM and the best match found. If this best match produced a higher match score than the gapped version of the TFBS, then it would be assessed by comparing it with the score of the known TFBS as described above. If this led to any comment except one including "diverged", that comment would be taken into account in the later analysis.

It will be evident that this could lead to two contradictory comments about the same possible homologue TFBS, as it could cause the comment "Probably conserved" to be added even if a "Probably diverged" comment had already been produced by the procedure described in the previous paragraph. Either conclusion could be argued to be plausible depending on whether one believed that the alignment program had inserted the gap in exactly the right place or not. The existence of these two contradictory comments would later cause the analysis to generate a conclusion of "Uncertain", and an explanatory

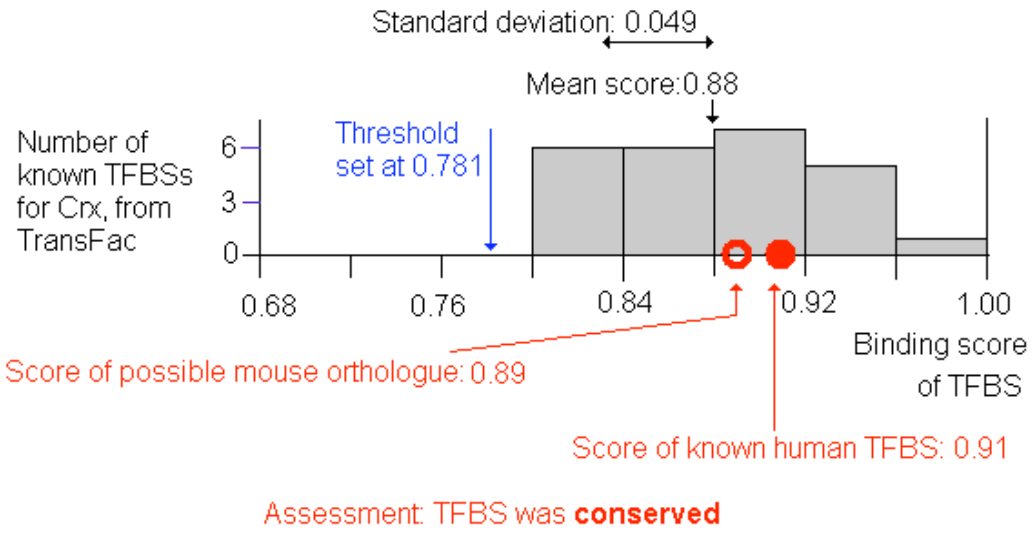


Fig 2.4 B (above). The possible mouse orthologue has a binding score which is above the threshold *and* almost as high as the score of the human TFBS (within two standard deviations). Thus it meets neither criteria for a non-conserved TFBS, and is assessed as *conserved*.

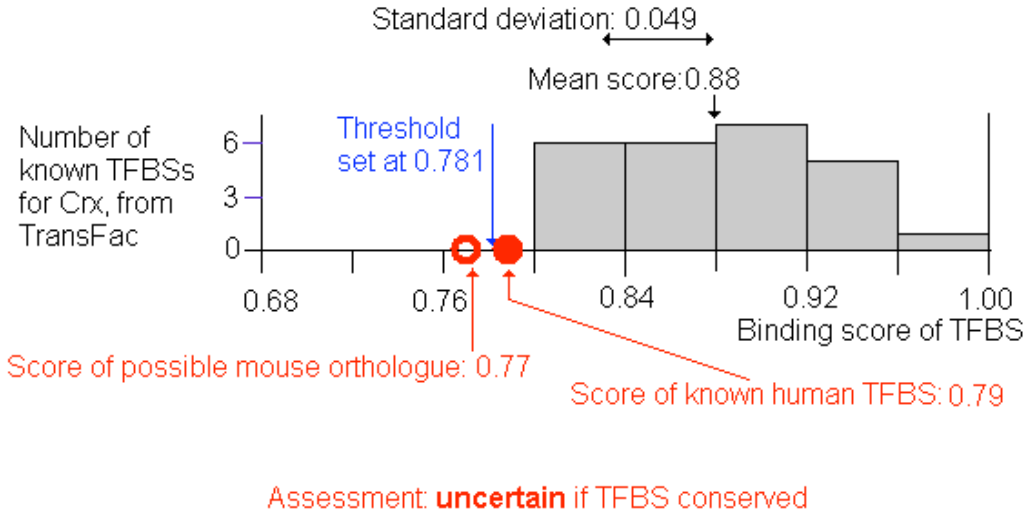


Fig 2.4 C (above). The possible mouse orthologue has a binding score which is below the threshold, but is almost as high as the score of the human TFBS (within two standard deviations). Thus it meets one but not both criteria for a non-conserved TFBS, and is assessed as *uncertain* whether the TFBS was conserved.

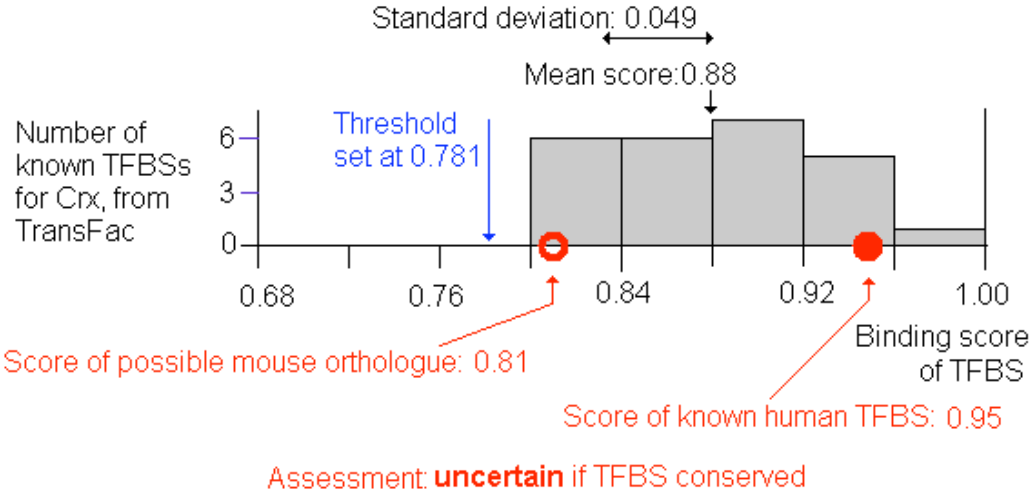


Fig 2.4 D (above). The possible mouse orthologue has a binding score which is above the threshold, yet well below the score of the human TFBS (more than two standard deviations below). Thus it meets one but not both criteria for a non-conserved TFBS, and is assessed as *uncertain* whether the TFBS was conserved.

remark in the output "Mixed evidence, could be interpreted as conserved or diverged".

The original motive for developing the above procedure was that manual examination occasionally revealed cases where this seemed necessary (eg, as in fig 2.8). Now, a recent publication (Pollard et al., 2006) helps to retrospectively justify it. This publication quantified how often misaligments occur; it showed that, when a TFBS is misaligned, often the misalignment is only by a few bases. This implies that reputable alignment programs will sometimes insert a small gap in the wrong place. Thus, a gap against a TFBS might be an incorrect alignment, and so it is prudent to consider whether a re-arranged version of the alignment might lead to an alternative conclusion. In effect, the procedure described above does this.

The above applied when the possible *homologue* TFBS contained gaps. However, when the known TFBS itself contained gaps, then it was analysed using a procedure very similar to that for ungapped cases, except that: (i) the binding score of the known TFBS was, of course, based on the ungapped sequence of that TFBS; (ii) the "possible homologue TFBS" was too long, since it was aligned against a gapped version of the known TFBS, and so the best binding sequence within it was found by searching using the PWM, and used in subsequent analysis.

Noting Duplicates

If a possible TFBS homologue had nearly the same gene symbol, the same species, and the same DNA sequence as a known TFBS already analysed earlier in the run, then the software generated a comment stating that it "Duplicates earlier site".

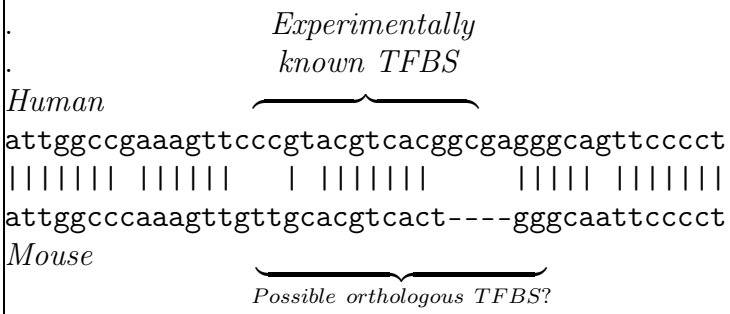
Examination of "neighbourhood" (DNA flanking a TFBS)

The "neighbourhood" sequence was defined as the TFBS plus 50 bases on either side. 50 bases was an arbitrary choice. Years after that choice was

Table 2.8: Example illustrating alignment gaps against a TFBS

An actual example showing a human TFBS aligned against gaps in mouse sequence

From the TransFac database, the transcription factor CREB is known to bind the human genome approximately 70 bases upstream from the ADRB2 gene. Aligning that portion of the human genome against the corresponding portion of the mouse genome (using Blastz) gaves the following:



Is the TFBS conserved and present in the mouse? One possible interpretation would be that the TFBS was affected by a deletion event in the mouse lineage, and that this must have destroyed the TFBS. However, even if we assume the deletion event occurred, it is not clear if this will have destroyed the ability to bind CREB. The effect of the deletion has been to delete the last two bases of the TFBS (gc in human) which were, in effect, replaced by the two bases immediately to the right of the deletion (gg in this case). Therefore the effect of the deletion on the TFBS sequence was only to change the last base from c to g. The PWM for CREB suggests this will have little or no effect on binding, since both c and g are often found as the last base in a CREB TFBS. This suggests that it is possible that the TFBS survived the deletion event.

It will be noted that in the procedure just described, in order to decide if CREB could bind the mouse genome at this location, the gaps were left out, and the ability to bind the gap-free sequence was assessed.

In the case just given, the two bases to the right of the sequence was considered as part of the possible TFBS (even though they were not aligned with the known human TFBS). In this case it is clear that sequence *to the right* should be considered. But in other cases, if the gap was near the centre of the TFBS, it might not be obvious whether sequence to the left or to the right should be used, making it necessary to try both.

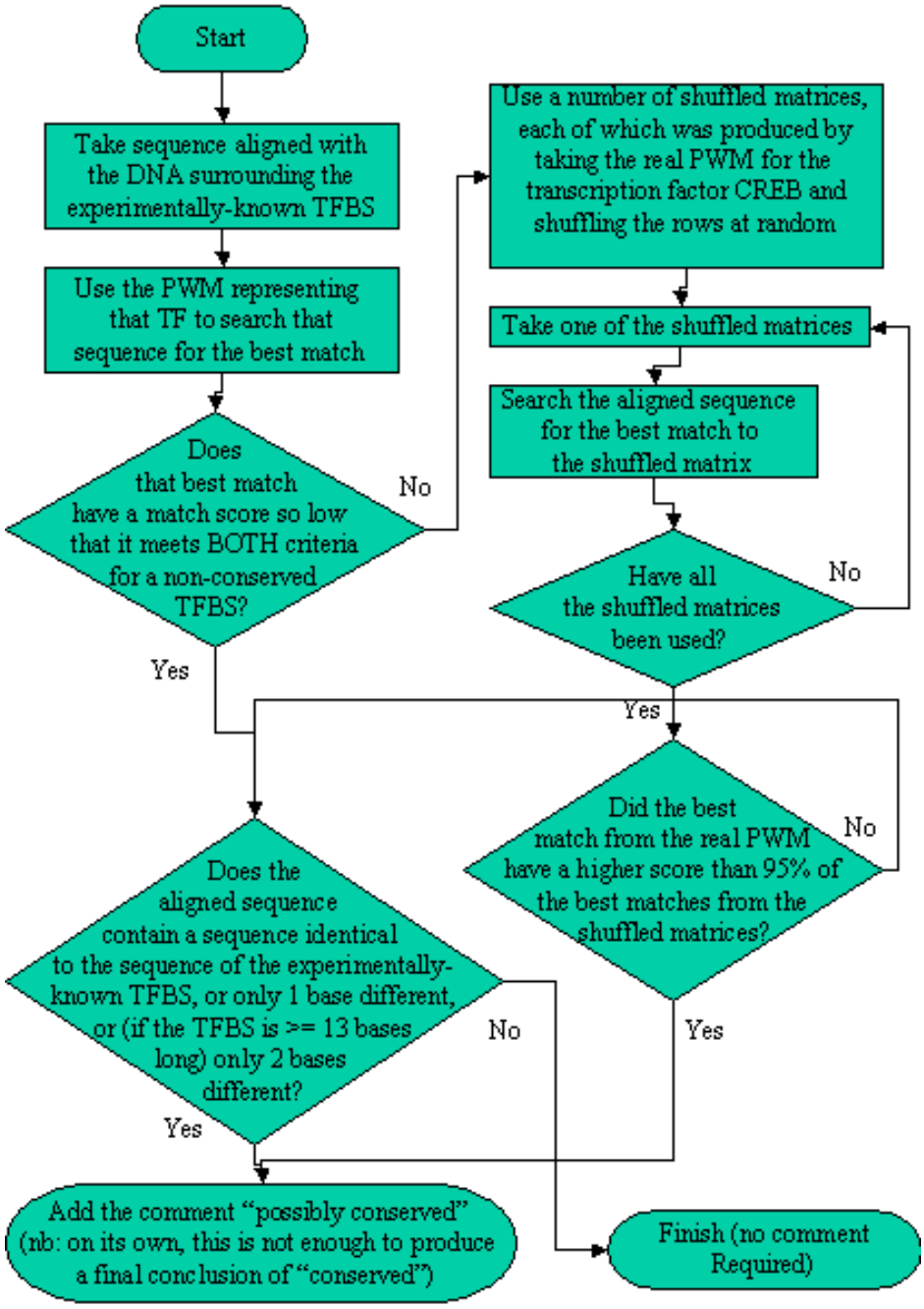
made, a publication reported that mammalian regulatory elements occur in conserved blocks of sequence which are 96 or 112 bases wide (median value) (Xie et al., 2007); hence 50 bases either side appears to be quite appropriate.

The neighbourhood sequences of the two species were compared to see how well conserved they were, the measure being the percentage identity. For this purpose, the TFBS itself was left out. 50 bases on either side were used, even if that meant using genomic sequence that was outside the aligned region determined by BlastZ. Obviously the sequences had to be aligned before measuring percentage identical. Since some sequence was not alignable by Blast2 (which was being used as the main alignment program at the time this module was written), the sequences were aligned by T-Coffee (Notredame et al., 2000). In early work, the percentage identity used was from a calculation in which gaps were counted as mismatches; results using this will be presented in table 2.17, and results in chapter 3 also used this measure (except for the most recent results, which are marked chip-chip run H3). For later work, percentage identity was used from a calculation in which gaps were ignored.

Statistics on the GC content of the neighbourhood were also produced.

The neighbourhood sequence was searched for any high-scoring matches to the PWM. This was in case the alignment procedure had made a small mistake in aligning the sequences, so that the known TFBS was aligned against a sequence a short distance from its true orthologue; the sequence it was aligned against would very likely have a low binding score, making the TFBS would appear to be non-conserved, even if it was in fact conserved. The true orthologue of the TFBS might be present as a "match" with a high binding score in the neighbourhood sequence of the comparison species; thus, it was thought desirable to detect these cases, and classify them as "Uncertain" if they would otherwise have been classified as non-conserved. Two years after implementing this procedure, the justification for it was strengthened by a publication (Pollard et al., 2006) which showed that incorrect alignments can occur, but that when this happens, in a high proportion of cases a conserved TFBS is only a few bases out of alignment with its orthologue TFBS. The procedure just outlined addresses this problem.

Figure 2.5: Flowchart showing how "neighbourhood" was examined
Flowchart showing how "neighbourhood" sequence (that is, sequence within 50 bases of a possible homologue to a TFBS) was searched. The aim was to deal with cases where a TFBS was in fact conserved but the alignment was slightly incorrect. (The two criteria for a non-conserved TFBS were detailed on page 94).



This procedure is described in more detail in figure 2.5. The sequence searched was that returned from BlastZ (that is, not included sequence that was not aligned), after removing gaps. Because a match with a moderately high binding score could easily occur by chance in the 100 bases of neighbourhood, this could cause a problem by giving the appearance of a conserved TFBS even if the TFBS was, in fact, not conserved; so this was guarded against by the shuffled-matrix search detailed in fig 2.5.

Whilst this procedure might cause a "Possibly conserved" comment, that would never be regarded as being sufficient evidence for producing a final conclusion that a TFBS was conserved. It could, however, cause a final conclusion of "Uncertain" to be given when a TFBS might otherwise have been classifed as "non-conserved".

Awkward alignments

As noted earlier (page 93), if the known TFBS was between two aligned regions, then a special procedure was followed. A simple illustration of the situation is shown in Table 2.9, which explains why the entire "between-aligns" sequence has to be searched before reaching any definite conclusion that the TFBS was conserved or non-conserved. The actual procedure used is shown in fig 2.6. There is a considerable risk of getting a high-scoring match by chance when searching a sequence which could easily be hundreds of bases long; so to prevent this causing misleading results, the shuffled-matrix search detailed in fig 2.6 was employed.

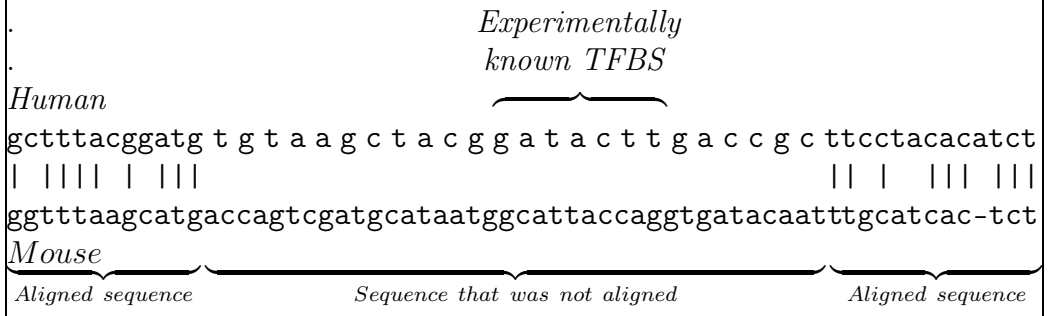
Generating a conclusion

It will have been noticed from the preceding text that a "comment" might be generated at various points during the anlaysis. This could result in several comments about a single case, which were not always consistent. For instance, comparing the known TFBS against the sequence to which it was aligned could generate a comment suggesting the TFBS was "diverged", whilst if the nearby sequence from the comparison species contained a high-scoring

Table 2.9: Simple example illustrating when a TFBS is "between-aligns"

An imaginary example showing a "between-aligns" situation.

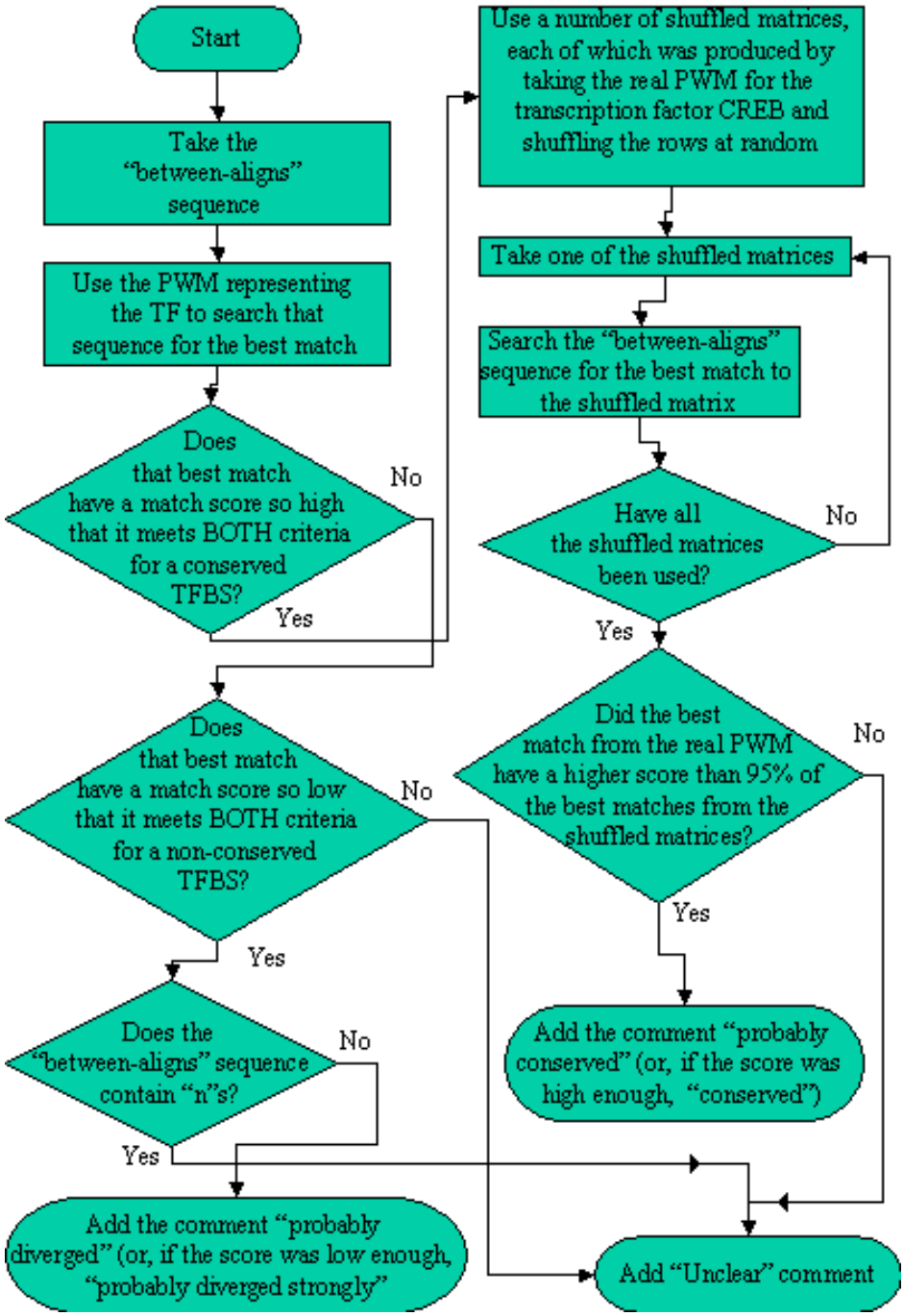
Human and mouse sequence has been submitted to an alignment program, which has found two aligned portions. Between these two portions is sequence for which no alignment could be detected, as shown below:



The sequence between the two aligned portions is referred to here as "between-aligns" sequence.

There is a TFBS in the human "between-aligns" sequence. Could this TFBS be conserved in the mouse? If it is conserved, presumably it is present somewhere in the mouse "between-aligns" sequence. Using the alignment information alone, we cannot tell exactly where in the mouse it would be (especially since, in this example, the "between-aligns" sequence is longer in the mouse than in the human). Thus, the orthologous TFBS in mouse (if it exists at all) has to be located by searching the entire mouse "between-aligns" sequence for a good match to the expected TFBS sequence. The details of this procedure are given in figure 2.6.

Figure 2.6: Flowchart showing "between-aligns" sequence analysis Flowchart showing "between-aligns" sequence analysis. The TF CREB is used as an example. (The two criteria for a non-conserved TFBS were detailed on page 94).



match to the PWM, that could generate a "Possibly conserved" comment; either of these two contradictory comments might be correct, the former being sensible if the alignment was correct, but the latter being plausible if a minor misalignment had occurred.

In addition, where there was some uncertainty about the exact position of the known TFBS (see "Locating the precise TFBS", page 87) analysis of the 2nd-best site would have produced additional comment(s).

To summarise all these comments into a single conclusion, the procedure in fig 2.7 was followed. This procedure also generated a explanatory remark (which might be blank if not needed).

This single conclusion was the one used in the conserved/uncertain/diverged classifications reported in the Results section.

The basic idea behind this procedure was that any comments indicating a problem, or any contradictory comments, should lead to the conclusion "Uncertain". Only if the comment(s) consistently indicated a conserved TFBS would the conclusion "Probably conserved" be reached; and similarly for the diverged TFBSs.

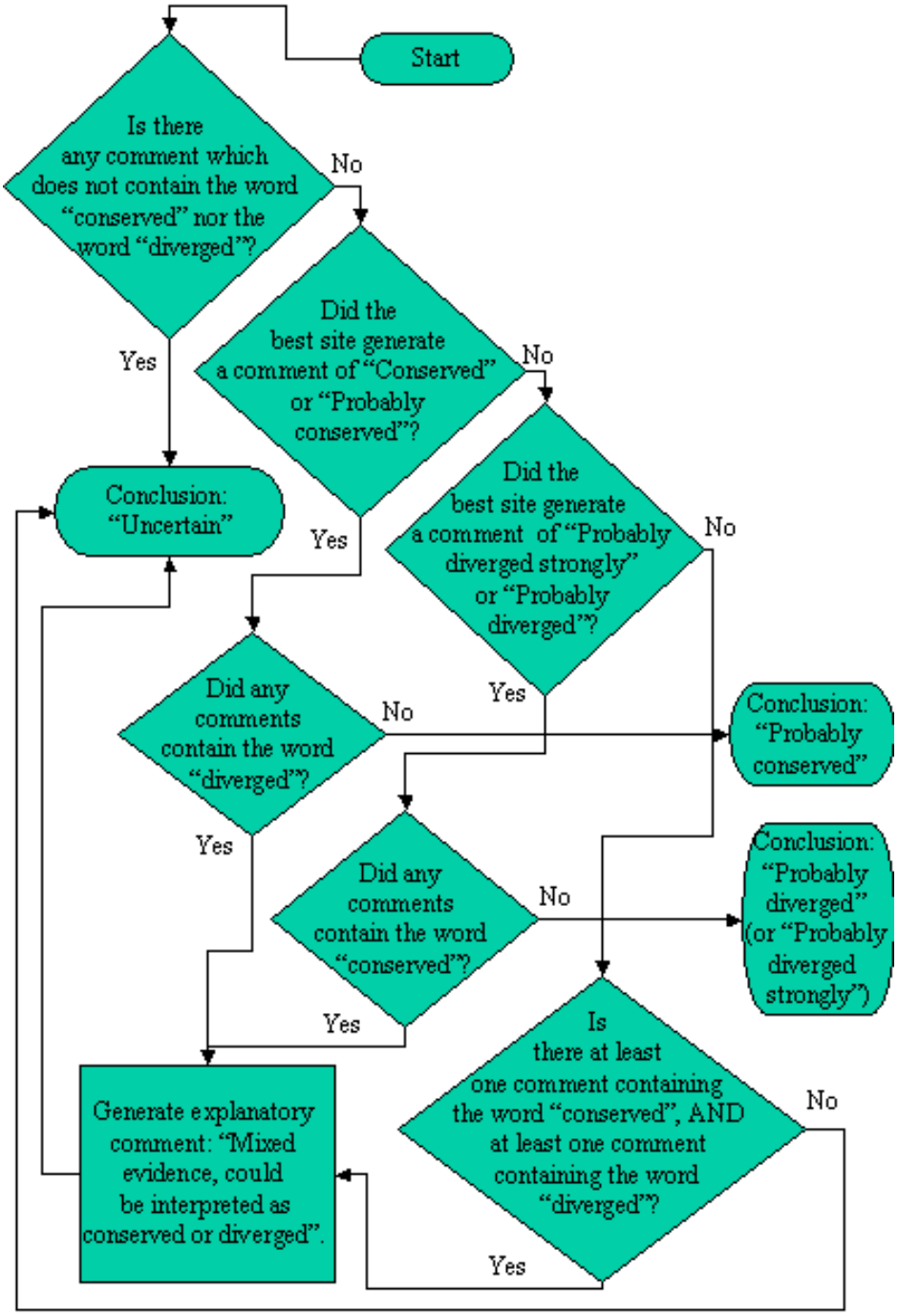
2.2.5 Distinguishing versions of outputs

Each time the analysis program was run, the results from that run were given an identity code, such as "Run J". These will be referred to here, so it is clear when two results are derived from the same dataset. Later letters refer to later versions, so "Run A" is early, whilst "Run Q" is the most recent version reported in this thesis. Analysis of the Reserve List and chip-chip data were lettered in separate series.

2.2.6 Minor utilities

To allow the user to control the amount of information put in the output file, on starting the software the user was asked how verbose an output they

Figure 2.7: Flowchart showing how to obtain an overall conclusion
Flowchart showing how to obtain an overall conclusion for the comparison of a known TFBS with DNA from another species. Comment(s) from previous stages of the analysis must be available at the start.



Declaration: no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning, except for the reuse of some Perl modules (altered as required), originally written during the candidate's Master in BioInformatics course (University of Exeter)